# On Provable Copyright Protection for Generative Models

Year: 2023 | Citations: 109 | Authors: Nikhil Vyas, S. Kakade, B. Barak

---

## Abstract

There is a growing concern that learned conditional generative models may output samples that are substantially similar to some copyrighted data $C$ that was in their training set. We give a formal definition of $\textit{near access-freeness (NAF)}$ and prove bounds on the probability that a model satisfying this definition outputs a sample similar to $C$, even if $C$ is included in its training set. Roughly speaking, a generative model $p$ is $\textit{$k$-NAF}$ if for every potentially copyrighted data $C$, the output of $p$ diverges by at most $k$-bits from the output of a model $q$ that $\textit{did not access $C$ at all}$. We also give generative model learning algorithms, which efficiently modify the original generative model learning algorithm in a black box manner, that output generative models with strong bounds on the probability of sampling protected content. Furthermore, we provide promising experiments for both language (transformers) and image (diffusion) generative models, showing minimal degradation in output quality while ensuring strong protections against sampling protected content.